A PROBABILISTIC FRAMEWORK FOR OBJECT RECOGNITION IN VIDEO

Omar Javed, Mubarak Shah

Computer Vision Lab
University of Central Florida
Orlando, Fl, 32816

Dorin Comaniciu

Real-time Vision & Modeling Department
Siemens Corporate Research
Princeton, NJ 08540

ABSTRACT

We propose a solution to the problem of object recognition given a continuous video sequence containing multiple views of an object. Initially, object models are acquired from images of the objects taken from different views. Recognition is achieved from the video sequences by employing a multiple hypothesis approach. Appearance similarity, and pose transition smoothness constraints are used to estimate the probability of the measurement being generated from a certain model hypothesis at each time instant. A smooth gradient direction feature that is quasi-invariant to illumination changes and noise is used to represent the appearance of object. The pose of the object at each time instant is modelled as a von Mises-Fisher distribution. Recognition is achieved by choosing the hypothesis set that has accumulated the maximum evidence at the end of the sequence. We have performed detailed experiments demonstrating the viability of the proposed approach.

1. INTRODUCTION

Object recognition is an important task with applications in the areas of automated surveillance, human-computer interfaces and video retrieval. Although a large number of object recognition methods have been proposed, a majority of these approaches are based on matching a single image with object models. The advantage of using video instead of a single image for recognition is that not only more than one pose of the object can be visible but there is also a smoothness in pose transitions which can be exploited for recognition.

In our proposed method, the measurement (representing appearance of object) from each frame of the test sequence is matched to the stored models. We use the gradient direction field to represent the appearance of the object. A spline based technique is used to estimate the gradient directions. The probability that the measurement is generated from a certain object hypothesis depends on both the similarity of the measurement to the model and the change of the estimated pose over time. The appearance similarity is computed by the probabilistic matching of the observed object's gradient features to the models. The pose transition probability is obtained by assuming that the pose angles at time t are distributed as a von Mises-Fisher distribution with the mean parameter equal to the pose angles at time $t-1$. A maximum a posteriori (MAP) estimation frame work is used to obtain the hypothesis set that has accumulated the maximum evidence for recognition.

The details of the related work are given in Section 2. Modelling and feature extraction of the objects are discussed in Section 3. The probabilistic formulation of our video based recognition problem is discussed in Section 4. An efficient method to obtain the MAP estimate is given in Section 5. Results are given in Section 6.

2. RELATED WORK

In order to develop models of objects for recognition, Seibert and Waxman [7] automatically cluster frames from training sequences into view categories called aspects. These aspects, as well as aspect transitions, are learned from the training sequences and are stored in the form of an aspect graph. Recognition is carried out by matching corner like features of an object to the stored aspects. The matching score depends on both the current aspect as well as on the history of aspect transitions. Rao [6] uses a robust form of Kalman filter to generalize the PCA based recognition methods over sequences. The filter learns the dynamic model of the view transitions of objects from the training sequences. The method assumes that the change in pose of the object (or camera view) in the test videos will be similar to the training videos. Our work imposes the less restrictive condition that the pose of the object should change smoothly.

Zhou et. al [9] propose a method for face tracking and recognition simultaneously using particle filters. A first order motion model is used for tracking. In addition, an adaptive appearance model is used to handle inter-frame appearance changes of the object. One assumption of the method is that the stored models consists of only frontal face images. Li et. al [3] build a model of each object consisting of discretely sampled views of the object from the view sphere.

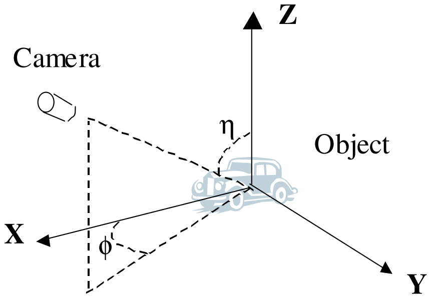


Fig. 1. The object's pose with respect to the camera in terms of two camera viewing angles (η,ϕ).

Edge maps of each view of the objects are used as features for matching. A generalized Hausdorff metric is used as a distance measure to identify both the object and its pose. At the end of the sequence, an object hypothesis with the least Hausdorff score is accepted, if its pose transitions are within a threshold. One difference in our approach is that we have a unified pose smoothness and model match measure rather than having two separate tests for object identification. Moreover, our single view matching depends on the regularized gradient direction obtained from image. Thus there is no loss of information as opposed to the thresholding in edge maps. Another salient feature of our method is the use of appropriate spherical distributions to model the angular pose data.

3. MODEL ACQUISITION AND FEATURE EXTRACTION

In this paper, models are constructed for recognition purposes by taking images of each object at various camera tilt angles, with the object rotating horizontally. Note that the pose of the object on a level ground can be characterized by two camera viewing angles (η,ϕ), using a object centered coordinate system (see Fig. 1).

Once images of the objects are obtained, the next issue is: what features should be used to represent the object. A large variety of features and their combinations have been used by previous recognition approaches including color, edges, gradients, oriented Gabor responses, wavelets, corners, lines etc. We want to use a feature that is invariant to illumination changes and is robust to noisy measurements. One feature that is not affected greatly by illumination changes is the gradient direction of the pixels in an image. However, local gradient direction is sensitive to noise. To overcome this problem, we represent the gradient direction field by splines. Splines impose a smoothness constraint over the gradient direction field. In addition, splines offer a compact representation of the field thus reducing the storage requirements for the object models. Note that splines have been successfully used by, Szeliski et. al [8] to estimate optical flow and by Le Pennec and Mallat [5] to compute horizontally or vertically parallel geometric flow.

Our aim is to find the direction at each pixel position along which the image gray levels have regular variation. The solution to this problem consists of minimizing the squared error

$$\sum_{x,y} \Big(f_x(x,y)\cos(\theta(x,y)) + f_y(x,y)\sin(\theta(x,y)) \Big)^2 \quad (1)$$

where f_x and f_y denote image derivatives. We represent the gradient direction field $\theta(x,y)$ as two dimensional splines controlled by q parameters, $\hat{\theta_1},..,\hat{\theta_q}$ that lie on a coarser spline control grid. The value of gradient direction at a pixel (x,y) can be written as

$$\theta(x,y) = linearSum_i\big(\hat{\theta}_i S_i(x,y)\big)$$

where $i = 1..q$, and $S_i(x,y)$ are the basis functions that have a finite support. Each S_i is centered on a control grid point and is a spatially shifted version of a function S. We have implemented S as a bilinear interpolation function, i.e. $S(x,y) = (1-|x|)(1-|y|)on[-1,1]^2$. Note that a weighted sum of angles can not obtained directly because of the wraparound at 2π. The $linearSum$ function does a pairwise interpolation between the unit vectors representing the angles to get the correct value. In order to obtain the gradient flow we need to minimize Equation 1 with respect to $\hat{\theta_i}$. We use a trust region method [1] for the nonlinear minimization. The method uses the finite difference derivatives of the error function to obtain the approximate Jacobian matrix. The spline parameter values are constrained between 0 and 2π.

4. PROBABILISTIC FORMULATION FOR OBJECT RECOGNITION

Suppose we have models of n objects $O_1, O_2, \ldots, O_n$, each with q poses. Let $M_t = m_1, m_2, \ldots, m_t$ be the set of measurements till time t. Let $k_t^{i,p}$ be the hypothesis that measurement at time t belongs to the i^{th} object with pose p.

Now, a feasible solution $K_{t'}$ at time instant t' , where $K_{t'}$ is a set of hypothesis, is the one that satisfies the following constraints

- For all t, where $t = 1, 2, \ldots t'$, $\exists k_t^{i,p}$ such that $k_t^{i,p} \in K_{t'}$, i.e., each measurement is used in the solution $K_{t'}$.
- if $k_t^{i,p} \in K_{t'} \wedge k_u^{l,j} \in K_{t'}$ then $t \neq u$, i.e., only one hypothesis at each time instant belongs to solution $K_{t'}$.
- if $k_t^{i,p} \in K_{t'} \wedge k_u^{l,j} \in K_{t'}$ then $i = l$, i.e. all hypothesis in the solution $K_{t'}$ are for the same object.

The probability of a feasible solution $P\left(K_{t'}|M_{t'}\right)$ at time t', can be written as,

$$P(K_{t'}|M_{t'}) = \Big(\prod_{t=2...t'} P(k_t^{i,p}, k_{t-1}^{i,j}|m_t)\Big)P(k_1^{i,a}|m_1),$$

by making the assumption that the current hypothesis only depends on the previous hypothesis and the current measurement.

Now, using the Bayes rule on the product term,

$$P(k_t^{i,p}, k_{t-1}^{i,j}|m_t) = \frac{1}{c}P(m_t|k_t^{i,p}, k_{t-1}^{i,j})P(k_t^{i,p}|k_{t-1}^{i,j})P(k_{t-1}^{i,j})$$

Here, c is the normalization constant. The last term in the above given equation, i.e., the probability of a hypothesis, is assumed to be uniformly distributed. Now the solution of the recognition problem is a solution K', in the solution space ω, that maximizes the posterior,

$$\begin{aligned} K' &= \arg\max_{K \in \omega} \Bigg(\Big(\sum_{t=2...t'} log\big(P(m_t|k_t^{i,p}, k_{t-1}^{i,j}) \\ &\qquad P(k_t^{i,p}|k_{t-1}^{i,j})\big)\Big) + log(P(m_1|k_1^{i,a}))\Bigg) \qquad (2) \end{aligned}$$

The first term in the above equation is the probability of obtaining a certain measurement given object identity and pose, i.e., the measurement to model match probability. The second term is the probability of the current pose of an object given the previous pose. In order to maximize the posterior we need to estimate the pose transition and appearance matching probabilities. This issue is discussed in the following.

4.1. Estimation of Pose Transition Probabilities

The pose of an object should change smoothly if the motion of the object (or the camera) is continuous. Thus there should not be large changes in the pose angles over subsequent measurements. Since we are working with spherical data, our probability model needs to handle wrap around of angles at 2π. We use the von Mises-Fisher (VMF) distribution [4] for a 3 dimensional sphere to model the transition probability $P(k_t^{i,p}|k_{t-1}^{i,j})$. The VMF distribution is unimodal with the mode at the mean direction. Moreover the VMF distribution is rotationally symmetric about the mean direction and has a finite support.

Suppose the pose j of a hypothesis at $t-1$ is represented by angles (α, β). Assuming the pose at $t-1$ to be the mean of the VMF distribution, the probability of the current pose p, represented by angles (η, ϕ) is given as

$$f(\eta, \phi|\alpha, \beta) = \frac{\kappa}{2\sinh\kappa} e^{\kappa(\cos\eta\cos\alpha + \sin\eta\sin\alpha\cos(\phi-\beta))}\sin\eta$$

where $\kappa \geq 0$ is the concentration parameter and gives the spread of the distribution around the mean direction.

4.2. Estimation of Measurement to Model Matching Probabilities

For each incoming frame the measurement, m_t consists of the spline parameter vector $\hat{\boldsymbol{\theta}} = [\hat{\theta_1}, .., \hat{\theta_q}]$ that describes the gradient direction field of the observed object. Each component of this vector represents the gradient direction angle and has a range between 0 and 2π. We need to match the measurement to the stored object models. Note that object models are also represented by spline parameters. We assume that each component of the model parameter vector has a von Mises distribution. The von Mises distribution is the circular analog of the Gaussian distribution on a line [4]. The probability that the measurement is similar to the model i with pose p is given by

$$f(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\mu}}^{i,p}) = \prod_{z=1..q}\big(\frac{1}{2\pi I_o(\kappa)}e^{\kappa\cos(\hat{\theta_z} - \hat{\mu}_z^{i,p})}\big)$$

where $\kappa \geq 0$ is the concentration parameter and I_o is the Bessel function of the first kind. $\hat{\mu} = [\hat{\mu_1}, .., \hat{\mu_q}]$ is the spline feature vector of the model. Note that we have made the simplifying assumption that each component of the vector is independent of the other. We also require that the measurements and model vectors have the same dimension. Now that we have the probabilities for the pose transitions and appearance similarity, we can find the MAP object recognition estimate.

5. OBTAINING THE MAP ESTIMATE

The problem of finding a hypothesis set that maximizes the a posteriori probability can be modelled as follows: We construct a directed graph such that each hypothesis k is represented by a vertex. Each vertex representing a hypothesis at time $t-1$ is connected by a directed arc to all vertices representing hypotheses for the same object at time t. The weight of the arc joining vertices representing $k_{t-1}^{i,j}$ and $k_t^{i,p}$ is calculated as $-log(P(m_t|k_t^{i,p}, k_{t-1}^{i,j})P(k_t^{i,p}|k_{t-1}^{i,j}))$,(see Eq 2). The negative sign is used to convert the problem of finding the best hypothesis set into a minimization problem. The vertices representing the hypotheses at the first time instant are connected to a source vertex. The weight of the arc joining the source to the vertex hypothesis $k_1^{i,a}$ is calculated as $-log(P(m_1|k_1^{i,a})$. The vertices representing the hypothesis at time t' are designated as terminal vertices. The MAP solution is the set of vertices (hypotheses) lying on the minimum weight path between source and terminal vertices of the directed graph. We have used the Dijkstra's single source shortest path algorithm [2] to find the least weight path and therefore the object model that has accumulated the maximum evidence.

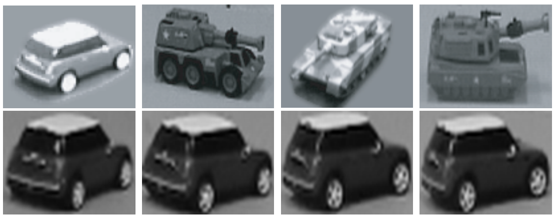


Fig. 2. The top row shows sample images of models 1 to 4 (from left to right) in the model database. The bottom row shows an object (same as model 1) extracted from a test sequence. The sequence was acquired using a hand held camera.

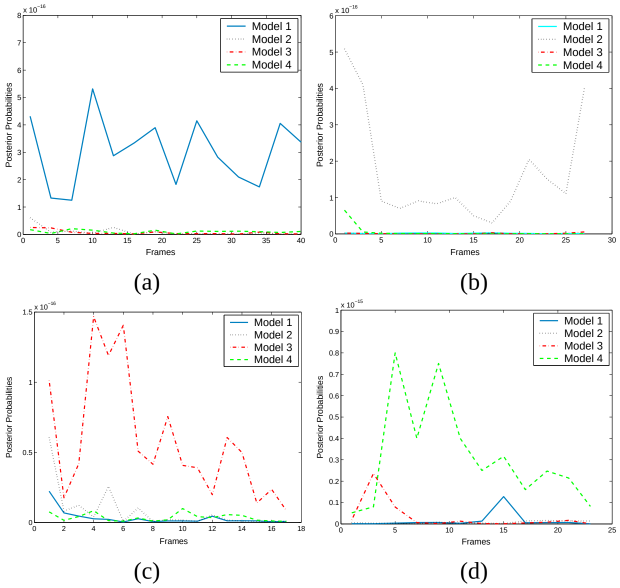
6. RESULTS

In order to evaluate the proposed approach we first obtained models of four different objects, by taking images of each object at different poses and computing the gradient direction field. A range of poses were obtained by varying ϕ by $3°$ in $[0°, 360°]$ and ρ taking two values $60°$ and $45°$.

Four sequences were taken to test the recognition algorithm. In the test sequences background subtraction was used to delineate the objects. The delineated objects were scaled to a fixed size. The Fig. 2 bottom row shows some images from the first test sequence. Note that the illumination in the test sequence is quite different from the model images. The results for the first sequence are shown in Fig. 3(a). The graph shows the posterior probability of the winning hypothesis set at each time instant and also the competing hypothesis sets representing the other three objects. The correct hypothesis set is clearly distinguishable from its competitors for the first sequence. Fig. 3 (b) and (c) show the results for the second and third sequence respectively. Correct results are obtained in both cases. In the fourth sequence, the wrong hypothesis set, representing model 3, has a larger probability than the correct hypothesis set representing model 4,for the first couple of frames as shown in Fig. 3(d). This is because models 3 and 4 (Fig. 2) are similar in appearance from the rear pose. However, as more evidence accumulates, i.e. other poses of the observed object become visible in the test sequence, the correct hypothesis set gets more probable. This clearly demonstrates the advantage of using video instead of a single image for recognition, since decisions are based on evidence accumulated over multiple frames.

7. CONCLUSION

In this paper, our goal was to incorporate the temporal information present in video for object recognition in a principled manner. We have presented a multiple hypothesis approach that uses appearance similarity and pose continuity constraints to come up with the best model hypothesis for recognition. In addition, we used directional distributions to model the appearance and pose data. For future work, we plan to explicitly incorporate the scale transition information along with pose changes for recognition.



(a) (b)

(c) (d)

Fig. 3. (a,b,c,d) Results for Sequences 1 to 4. The graph shows posterior probabilities of the hypothesis sets with the highest evidence for each object in the database.

8. REFERENCES

[1] M. A. Bracnh, T. F. Coleman, and Y. Li. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing*, 21(1), 1999.

[2] G. Brassard and P. Bratley. *Fundamentals of Algorithms*. Prentice Hall, 1996.

[3] B. Li, R. Chellappa, Q. Zheng, and S. Z. Ser. Model based temporal object verification using video. *IEEE Transactions on Image Processing*, 10(6), June 2001.

[4] K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley, 2000.

[5] E. L. Pennec and S. Mallat. Sparse geometric image representation with bandelets. *to appear in IEEE Trans. on Signal Processing*.

[6] R. P. N. Rao. Dynamic appearance based recognition. In *Proc. of CVPR*, 1997.

[7] M. Seibert and A. M. Waxman. Adaptive 3-d object recognition from multiple views. *IEEE Trans. on PAMI*, 14(2), February 1992.

[8] R. Szeliski and J. Coughlan. Spline based image registration. *International Journal of Computer Vision*, 22.

[9] S. Zhou, R. Chellapa, and B. Moghadam. Adaptive visual tracking and recognition using particle filters. In *Proc. of ICME*, 2003.